Problem Set 3 (51 Points)

6.7930/HST.956 Machine Learning for Healthcare

March 14, 2025

Submission Guidelines

Due 4/8/2025, 23:59 on Gradescope

Please make a submission only if you are registered as a regular student. Submit your write-up as `[mit_email]_pset1.pdf` (*e.g.,* `demirel_pset1.pdf`), with all written work in a single PDF file following the problem set structure. Please append your code at the end of the report. For those who prefer to typeset in LaTeX, the source code of this file is here .

Submit your own work and note any collaborators - if none, state "Collaborators: none." If using external sources, cite them properly. You should be able to explain your solutions verbally. Please do not share your code or report with anyone inside or outside of the class, nor post them publicly online. Course staff welcomes any questions about these policies.

1 Problem 1: Identifying Causal Elements (10 points)

Below we present two scenarios. For each scenario, please explain whether or not the problem is "causal" in nature. If so, describe the relevant covariates (X), treatments (T), outcomes (Y), and any unobserved confounders (U) that might pose problems.

1.1 (5 points)

You have collected physiological time series data from patients during surgical procedures controlled by the continuous administration of anesthetic drug D at various dose levels. The collected data includes EEG traces, heart rate, SpO2, and blood pressure measurements. While most patients survive these procedures, there remains a minor risk associated with them. Your goal is to optimize the protocol for controlling drug dosage levels over time to maximize patient survival rates.

1.2 (5 points)

You work at a hospital and have just been assigned to help a new payer (*e.g.*, insurance agent) fulfill a unique request. This payer wishes to receive advance estimates of billing codes that will eventually be assigned to their patients, specifically requesting predictions of each patient's final billing codes after only their first 24 hours of care. Contractually, the payer is prohibited from using these estimates to influence their acceptance or dispute of claims; rather, they use this information to proactively assess their revenue streams. To accommodate this request, you plan to develop a predictive model based on retrospective data that will analyze the first 24 hours of a patient's stay to forecast the final ICD-10 codes they will receive at discharge.

2 Problem 3: Analyzing the Impact of Surgical Fellowship on Operating Room Performance (15 points)

- Z : Whether the surgeon graduated from Harvard Medical School (1 if yes, 0 if no)
- X : Whether the surgeon completed > 100 successful surgeries in 2015 (1 if yes, 0 if no)
- T : Whether the surgeon was selected for the surgery fellowship program during 2016-2017 (1 if yes, 0 if no)
- Y : Whether the surgeon completed > 200 successful surgeries in 2018 (1 if yes, 0 if no)
- W : Whether the surgeon cumulatively completed > 500 successful surgeries in their lifetime by 2019 (1 if yes, 0 if no)

Assume the following:

- The number of successful surgeries in 2015 (X) depends solely on surgeon's prior education (Z).
- Selection to the surgery fellowship program (T) depends on surgeon's prior education (Z) and other unmeasured factors not included in this model.
- The number of successful surgeries in 2018 (Y) depends on fellowship participation (T), prior education (Z), and the number of successful surgeries performed in 2015 (X).
- The cumulative number of successful surgeries by 2019 (W) depends on the surgeon's performance in 2015 (X) and 2018 (Y).

2.1 (5 points)

Draw the causal graph that represents this setting. Explain the connections between vertices.

2.2 (10 points)

The following is the data collected for the experiment.

Patient ID	Z	X	T	Y	W
1	0	1	0	0	1
2	0	0	0	0	0
3	0	1	1	1	1
4	0	0	0	1	0
5	1	1	1	0	0
6	1	0	1	1	1
7	1	0	1	1	1
8	1	0	0	0	1

Table 1: Data for experiment.

Calculate the Average Treatment Effect (ATE) of the fellowship (T) on 2007 success (Y), which is defined as $\mathbb{E}[Y(1) - Y(0)]$ where $Y(0)$ and $Y(1)$ denote the *potential* (counterfactual) outcomes under different treatment options, $T = 0$ and $T = 1$, respectively. Use covariate adjustment and empirically estimate the probabilities/expectations from the data. Clearly show your calculations and explain the steps.

3 Survival Analysis (10 points)

In this section, you will analyze data from a 24-month RCT to compare survival outcomes between a treatment group and a control group.

Below is a simplified dataset of 20 patients with their survival information:

Patient ID	Group	Survival Time (months, t_i)	Event (1=death, 0=censored)
1	Treatment	24	0
2	Treatment	18	1
3	Treatment	15	0
4	Treatment	13	1
5	Treatment	11	0
6	Treatment	10	1
7	Treatment	9	0
8	Treatment	7	1
9	Treatment	6	0
10	Treatment	4	1
11	Control	20	1
12	Control	17	1
13	Control	14	0
14	Control	11	1
15	Control	9	1
16	Control	8	1
17	Control	6	1
18	Control	5	0
19	Control	3	1
20	Control	2	1

Table 2: Patient survival data for treatment and control groups

3.1 Part A: Computing Kaplan-Meier Survival Estimates (5 points)

Compute the Kaplan-Meier (KM) survival estimates for both treatment and control groups using the formula: $S(t) = \prod_{t_i \leq t}(1 - \frac{d_i}{n_i})$ Where d_i is the number of death events at time t_i and n_i is the number at risk at time t_i. You can use the following template for your calculations.

Table 3: Kaplan-Meier computation table template

Time t_i	n_i	d_i	c_i	$1 - \frac{d_i}{n_i}$	$S(t_i)$

Note that c_i in the table above is the number censor events at time t_i.

3.2 Part B: Visualization and Interpretation (5 points)

1. Plot the KM curves for both groups on the same graph based on your computations in Part A.
2. Calculate the log-rank test statistic to compare the two survival curves. First calculate the chi-square test statistic. Then compute a p-value for that statistic for the null hypothesis that states two survival curves are the same. While doing that, use a degree-of-freedom of one for the chi-squared distribution as there are two groups we are comparing. Finally when deciding on whether or not reject this null hypothesis, use a significance level of $\alpha = 0.05$ (see lecture slides , pages 20-22).
3. Comment on the effectiveness of the treatment based on your analysis.

4 Propensity score re-weighting and covariate adjustment for ATE estimation (16 points)

Here you will apply some of the causal inference tools you learned on a real-world dataset. Your main task is to estimate the average treatment effect (ATE) of quitting smoking (T) on weight gain (Y), using the NHEFS dataset .

Your main learning objectives include:

- Understand how confounders, when unadjusted, can introduce bias into the ATE estimate.
- Learn how to implement propensity score re-weighting to estimate the ATE in Python.
- Learn how to implement covariate adjustment strategies to estimate the conditional average treatment effect (CATE) as well as ATE in Python.

Both code and write-up questions can be found in the starter notebook . Please save the notebook into your drive (File $->$ Save a copy in Drive) before starting to work on it.

Implement your answers to the coding questions (To-Do parts) inside the notebook. Then, please answer 8 "Write-up" questions which can also be found inside the notebook. Append your final code at the end of this report.